Model Considerations for Memory-based Automatic Music Transcription

Štěpán Albrecht* and Václav Šmídl†

*University of West Bohemia, Czech Republic
†Institute of Information Theory and Automation, Czech Republic

Abstract. The problem of automatic music description is considered. The recorded music is modeled as a superposition of known sounds from a library weighted by unknown weights. Similar observation models are commonly used in statistics and machine learning. Many methods for estimation of the weights are available. These methods differ in the assumptions imposed on the weights. In Bayesian paradigm, these assumptions are typically expressed in the form of prior probability density function (pdf) on the weights. In this paper, commonly used assumptions about music signal are summarized and complemented by a new assumption. These assumptions are translated into pdfs and combined into a single prior density using combination of pdfs. Validity of the model is tested in simulation using synthetic data.

Keywords: Automatic music recognition, Stochastic modeling, parameter estimation
PACS: 43.60.-c, 43.60.Cg, 43.75.Xz, 43.75.Zz, 87.10.Mn

INTRODUCTION

Automatic music transcription (AMT) is a process of decomposing recorded music signal into a sequence of higher-level sound events. The entire AMT—i.e. resolving pitch, loudness, timing and instrument of all sound events in an input audio music signal [5]—is not theoretically possible [5], therefore practical AMT has to be restricted to a specific scenario. Commonly used scenarios are memory-based and data-based AMT. The former utilizes sound models corresponding to a certain musical instrument sound (allowing to identify the instruments), the latter utilizes only rules which hold in general. We are concerned with a special case of memory-based AMT. As another memory-based AMT system, that can be termed as the entire [5], Kashino's transcription system [9] is considered.

Intuitively, the problem can be understood as an 'inverse music sequencer', Fig. 1. Music sequencers have pre-recorded library of sounds (sound components) which are combined together to create music signal. An input to the sequencer is a MIDI file which contains information about beginning of music events in time, their duration, IDs of sounds (in our case the pre-recorded sound components), their amplitude and modification type. Component modification(s)—e.g. component truncation or pitch shifting—were designed to reduce the size of the pre-recorded library. In this paper we consider only component truncation as a possible modification. Output of the sequencer is the audio signal. The input of our 'inverse music sequencer' is the recorded music signal and the output is the estimated (transcribed) MIDI-like representation of music events.

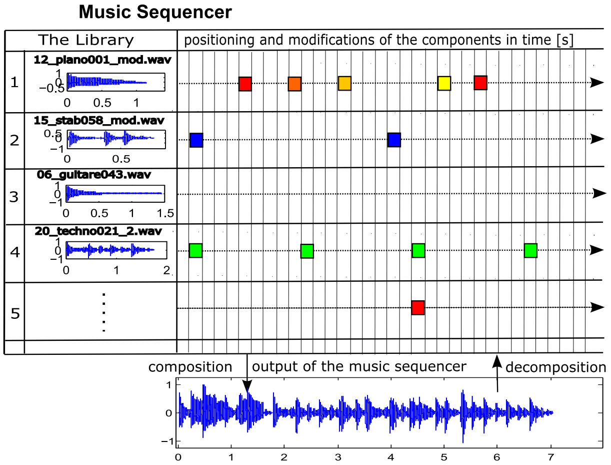

FIGURE 1. Music sequencer visualization: music events are depicted by squares. Their positions denote beginnings of the events in time. Each row represents one sound ID. Sound duration and modification is represented by different color.

In principle, the sequencer performs linear superposition:

$$y(t) \approx \sum_{k\in\mathscr{K}} \alpha(k,t)s(k,p_k), \tag{1}$$

where $y(t)$ is the ϕ-dimensional vector of measurements at time t composed of either time- or frequency-representation of the input music signal segment (frame); $\mathscr{K} \subset [1,\ldots,K]$ is a set of sounds *active* at time t (since we do not restrict the size of this set);the library of sounds S contains K sounds, each sound is formed by a sequence of L_k frames; k denotes the ID of the music event, position within the sound component (as requested by by component truncation) is denoted p_k, $1 \leq p_k \leq L_k$ and it is increasing with time, $s(k,p_k)$ denotes the p_kth frame in the kth sound component; $0 \leq \alpha(k,t) \leq 1$ denotes amplitude of the kth sound component at time t.

Model (1) is suitable representation of a sequencer, however, it is not suitable for the inverse operation since the number of possible configurations of the set $\mathscr{K}$ is enormous. Therefore, we re-parametrize (1) as being linear combination of all frames of all sound components, where amplitudes of the inactive frames are set to zero. The frames are indexed by a single index i, $i = 1,\ldots,N$, which denotes absolute position of a frame in the sound library:

$$y(t) \approx \sum_{i=1}^{N} a(i,t)f(i) + e(t). \tag{2}$$

Here, due to the change in indexing, we have changed symbols of weights from α to a and of s to f. The symbols for frames can be uniquely transformed to each other, e.g. $s(k,p_k) = f(\sum_{\kappa=1}^{k} L_\kappa + p_k)$. However, this is not the case for a and α since dimension of a is much larger, $a(i,t)$ now denotes weight of the ith frame. The activity of component in α is transformed into non-zero weight of the corresponding frames in a. Hence, the strong restriction from model (1) of only one frame from a sound component being active at time t was relaxed. This relaxation has both advantages and disadvantages.

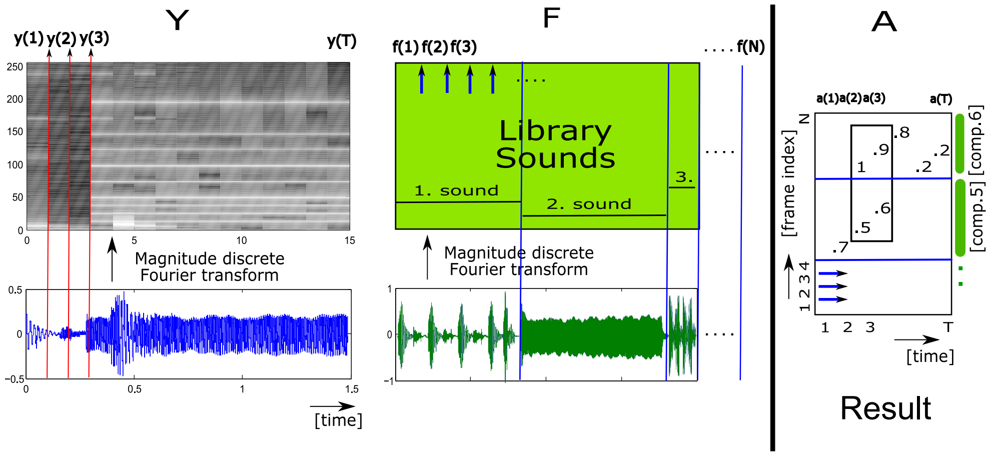

FIGURE 2. Example of an input signal composed of two overlapped library sounds. Amplitude matrix is flipped upside down.

The first advantage is that model (2) can be conveniently written in matrix form

$$Y = FA + E,$$

where matrices Y, F, A and E are composed as follows: $Y = [y(1), y(2), \ldots, y(T)]$, $F = [f(1), f(2), \ldots f(N)]$ $A = [a(0), a(1), \ldots a(T)]$, $a(t) = [a(1,t), a(2,t), \ldots, a(N,t)]$, $E = [e(1), e(2), \ldots, e(t)]$, as illustrated in Fig. 2. Second advantage is that model structure of model (2)—i.e. vector of measurements being linear combination of unknown parameters—is used in many statistical model for which there exist efficient parameter estimation methods. For example, linear regression, factor analysis [1], matching pursuit [11] and independent component analysis [2] (ICA) arise from (2) by imposing different assumptions on parameters A and F. These method are used in music processing, e.g. ICA for blind (unsupervised) source separation (BSS) techniques in monoaural input music signals [5].

The main disadvantage of the relaxation is that it allows to explain signal $y(t)$ by a combination of frames that are not valid from musical point of view. Since we want to avoid strict restrictions of model (1), we seek less strict representation of these assumptions. Since each of the methods mentioned above can be interpreted as a Bayesian estimator of A with different prior on A and F, we seek a smooth prior pdf on A that respects constraints of model (1) as close as possible. The challenge is to translate multiple pieces of prior knowledge about music signal into a single smooth pdf. We propose the following three-step approach: (i) each constraint is transformed a parametric pdf, (ii) these pdfs are joined into a single parametric pdf, and (iii) parameters of this density are estimated from artificially generated plausible realizations of A. The pieces are represented by a Gaussian pdfs for computational tractability.

The paper is organized as follows: common model consideration are presented in the first section; the main result—i.e. assessment of knowledge on musical signal and its composition into a pdf—is presented in the second section; the free parameters are trained on real data in the third section.

MODEL CONSIDERATIONS

The residues $e(\cdot)$ in model of observation (2) are assumed to have homogeneous Gaussian distribution, hence the whole sequence of observations can be written using matrix normal distribution [1]:

$$p(Y|A,F,\omega) = \mathscr{N}(FA, \omega^{-1} I_T \otimes I_\phi). \tag{3}$$

Here, $\mathscr{N}$ denotes normal distribution of matrix argument, ω is scalar precision parameter, I_T, I_ϕ denotes identity matrix of dimensions $T \times T$, $\phi \times \phi$, respectively.

The task is to estimate posterior density on matrix A, $p(A|F,Y)$. Following the Bayesian approach, we need to complement likelihood function (3) by prior density on ω and A, i.e. $p(\omega)$ and $p(A)$. For computational tractability reasons, we prefer to use priors conjugate to (3) which are Gaussian on A and Gamma on ω, [8]. Since omega is scalar parameter, its density is determined by two parameters of the Gamma density which may be chosen to yield flat prior, see e.g. [7]. However, $p(A)$ is parametrized by mean value of size NT and covariance matrix of size $NT \times NT$. Clearly a restricted parametrization of these parameters is required. Therefore, the constraints of music signal will be translated into structural properties of mean and covariance matrix of $p(A)$.

We consider the following phenomena that are specific for music signal:

(A) the amplitude matrix A is sparse. Note, from (1), that only less than $K \ll N$ elements of $a(t)$ should be non-zero.

(B) when the ith frame of a sound component is active in time t, and i is inside a sound component, the probability that the $i+1$ frame of the same component is active in time $t+1$ is high. No temporal correlation is assumed when i is at the end of a component.

(C) longer sequences of active frames are favored over shorter, on the other hand sequences should not exceed total length of the sound component, L_k. This assumption arise from the allowed truncation of the sound component.

(D) similarity measure between library frames with non-zero weight at time t should be low. Since frames within one sound component are similar, this assumption is a substitute for the constraint of only one active frame within a sound component from model (1).

Phenomena (A), (B) and (D) are commonly used in music signal processing. (A) is known as a sparsity constraint, see BSS methods in [5], (B) models temporal correlation of subsequent frames [5], [3], [4], (D) stems from statistical independence of sound sources, e.g. BSS in [5]. Phenomenon (C) is an extension proposed in this paper. These phenomena can be translated into Gaussian pdfs as follows.

Sparsity. Sparsity (A) is typically enforced by ‘pushing’ mean values toward zero mean, $a(t) \sim \mathscr{N}_{(A)}(\mu_{(A)}, \zeta^{-1} \cdot I_N)$, $\mu_{(A)} = 0$ with unknown scalar precision ζ. Note that due to restricted support of $a(i,t)$, the likelihood of $a(i,t) = 0$ can be further increased by choosing negative values of $\mu_{(A)}$.

Temporal dependence. Temporal covariance (B and C) of $a(t)$ with the previous vector realizations $a(t-1)$ is represented by mean value $\overline{a}(t-1)$ constructed as follows:

$$\begin{aligned}\overline{a}(t-1) \quad = \quad & [c, c+a(1,t-1), c+a(2,t-1), \ldots c+a(L_1-1,t-1), \\ & c, c+a(L_1+1,t-1), \ldots c+a(L_1+L_2-1,t-1), \\ & \ldots]'.\end{aligned} \tag{4}$$

Here, each row on the right-hand-side of (4) represents one sound component of length L_k. c denotes probability that a new component will start playing at any position – since we allow component truncation this constant is present in all elements of (4).

The preference for longer sequences (C) may be reinforced by suitable covariance structure design. We make the following choice for variance of the ith element of a sound component:

$$\lambda_i^2 = \left(\lambda + \gamma\left(1 - \left(\frac{p_k}{L_k}\right)\right)^2\right)^2 \tag{5}$$

For the purpose of this section, i is converted into k and p_k as used in model (1).

Note that the variance is linearly decreasing with index i, starting from $\lambda+\gamma$ at the beginning of a component, and reaching λ at the end of a component. The rationale is that we still prefer continuation of a sequence via its mean value. The exact steepness of the slope is modeled by parameter γ which will be identified from simulations.

Finally, the prior pdf for this two phenomena, (B) and (C) is:

$$p(a(t)|a(t-1)) = \mathscr{N}_{(BC)}(\overline{a}(t-1), \Lambda), \tag{6}$$

where Λ is a diagonal matrix with elements (5).

Similarity measure. The tones are supposed to be well isolated as in model (1). It is highly unlikely that very close frames from a sound component will be played together at the same time. This knowledge can be expressed by positive or negative covariance of the weights $a(i)$ and $a(j)$ of the ith and the jth sound in the library S. The pdf of this phenomenon is then: $p(a(t)) = \mathscr{N}_{(D)}\left(\mu_{(D)}, -\eta_1\Psi + \eta_2\right)$, where

$$\begin{aligned}\Psi(i,j) \quad &= \quad \cos\left(f(i), f(j)\right) \qquad &: i \neq j, \\ &= \quad \varepsilon \qquad &: i = j.\end{aligned} \tag{7}$$

This term is designed to penalize similar frames via negative correlation. However, we do not wish to encourage presence of any other tones via positive correlation. Since $\Psi(i,j) \to 1$ for $a(i,t)$ and $a(j,t)$ positive correlated, $\Psi(i,j) \to 0$ for $a(i,t)$ and $a(j,t)$ negative correlated, values of the Ψ must be rescaled. This is achieved by conversion $-\eta_1\Psi + \eta_2$. The diagonal of Ψ is composed of constants ε, chosen to be fixed since it has the same role as ζ in section (A).

In this Section, we make no assumptions on the mean values μ_D. This parameter will be optimized to have minimal information impact on the resulting pdf.

Remark 1:. Restriction of weights a to support $0 \leq a(i,t) \leq 1$ will be considered a posteriori.

ELICITATION OF PRIOR PDF

Each of the densities derived in the previous Section captures some important feature of the problem. The task now is to combine these densities into a single Gaussian pdf. We will use the standard geometric merging (also known as logarithmic pooling) [6]. The constructed pdf is build from pieces defined above as follows:

$$p(a(t)|a(t-1)) \propto \mathcal{N}_{(A)}(\cdot|\cdot)^{w_{(A)}}\mathcal{N}_{(BC)}(\cdot|\cdot)^{w_{(BC)}}\mathcal{N}_{(D)}(\cdot|\cdot)^{w_{(D)}} \tag{8}$$

where $\propto$ denotes equality up to a normalizing constant, $w_{(A)}, w_{(BC)}, w_{(D)}$ are scalar weights corresponding to each of the sources.

Without loss of generality, we will assume that these weights are equal to one. Note that each of the sources has covariance matrix multiplied by a scalar, for example ζ in $\mathcal{N}_{(A)}$. If $w_{(A)}$ was different from one, the weight can be absorbed by the covariance matrix as $w_{(A)}\zeta$. Thus, we consider that the weight of each pdf in (8) is already incorporated in their covariance structure.

It is easy to verify that geometric combination of Gaussian densities is again a Gaussian density:

$$\mathcal{N}(\bar{\mu}, \bar{\Sigma}) \propto \mathcal{N}(\mu_1, \Sigma_1)\mathcal{N}(\mu_2, \Sigma_2), \tag{9}$$

where $\bar{\Sigma} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$ and $\bar{\mu} = \bar{\Sigma}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$.

Application of formula (9) to pdfs $\mathcal{N}_{(A)}$ and $\mathcal{N}_{(BC)}$ yields $\mathcal{N}_{(ABC)}(\mu_{(ABC)}, \Sigma_{(ABC)})$, $\mu_{(ABC)} = \Sigma_{(ABC)}(\Lambda^{-1}\bar{a}(t-1))$, $\Sigma_{(ABC)} = (\zeta I_N + \Lambda^{-1})^{-1})$. Prior on phenomena (D) can be added similarly: $\tilde{\mu}_{all} = \tilde{\Sigma}_{all}(\Sigma_{(ABC)}^{-1}\mu_{(ABC)} + \Sigma_{(D)}^{-1}\mu_{(D)})$. However, since $\mu_{(D)}$ is not specified, this value can not be computed directly. Therefore, numeric value of μ_{all} is obtained by minimizing impact of unknown $\mu_{(D)}$ on the result of combination in the sense of Kullback-Leibler divergence [10]:

$$p(a(t)|a(t-1)) = \mathcal{N}(\mu_{all}, \Sigma_{all}) = \arg\min_{\tilde{\mu}_{all}, \tilde{\Sigma}_{all}, \mu_{(D)}} KL(\mathcal{N}(\tilde{\mu}_{all}, \tilde{\Sigma}_{all}), \mathcal{N}_{(ABC)}(\cdot,\cdot)). \tag{10}$$

Using the formula for Kullback-Leibler divergence of two Gaussian densities [8], it is easy to show that minimum of (10) is reached for $\mu_{all} = \mu_{(ABC)}$, $\Sigma_{all} = (\Sigma_{(ABC)}^{-1} + (-\eta_1\Psi + \eta_2)^{-1})$ (substitution of these values into (10) yields zero divergence which is a sufficient condition for minimum).

Elicitation of structure of the prior density $p(A)$ is now complete. The density is

$$p(A|\zeta, \lambda, \gamma, c, \eta_1, \eta_2) = \mathcal{N}_{(A)}(0, \zeta^{-1}I_N)\prod_{t=2}^{T}\mathcal{N}(\mu_{all}, \Sigma_{all}) \tag{11}$$

Here, the recursion is started at time $t = 1$ from $\mathcal{N}_{(A)}$. This choice is arbitrary. Potentially, nuisance parameters $\zeta, \lambda, \gamma, c, \eta_1, \eta_2$ can be treated as hyper-parameters and estimated from the observed data. However, for the purpose of verification of the developed prior structure, we will estimate their values form a training set of realizations of A using maximum likelihood approach. For convenience, we will maximize (11) using a numerical simplex method.

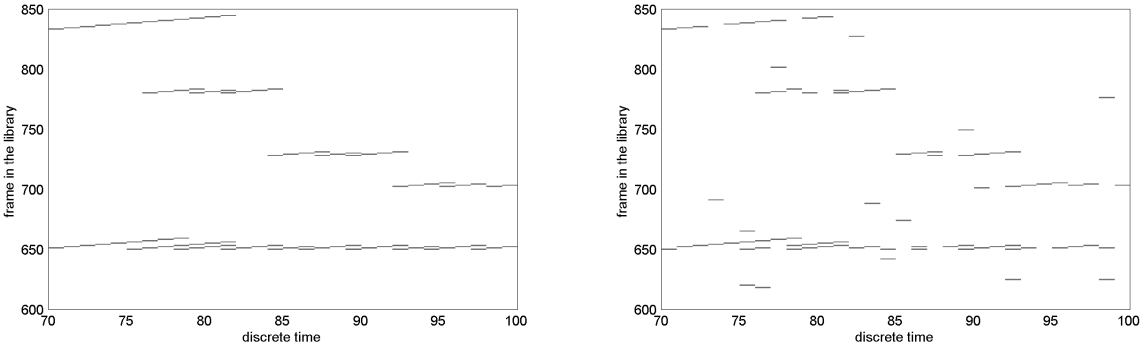

FIGURE 3. Amplitude matrix visualization. Left: amplitudes without corruption. Right: amplitudes corrupted by permutation.

EXPERIMENTAL VERIFICATION ON A SET OF MIDI FILES

Experimental verification was performed on data with the following parameters: number of frames in the library $N = 936$, number of sound components in the library $K = 35$, total length in time domain 2.5s. Each frame corresponds to 96ms in real time (corresponding to a segment of length 4096 when sample rate is 44.1 kHz.) The sound components ($s(k,t)$ and $f(i)$) were created using 'fm' synthesizer producing complex harmonic tones, their amplitude envelope was selected to resemble to a piano tone.

Two sets of music files were generated from MIDI of Mozart, Debussy, Beethoven and Chopin, the training set (50s) and the test set (20s). The training set was used to estimate nuisance parameters, while the test set was used to verify performance of the model.

The key criteria of performance is separation of valid music events representation from corrupted one. Hence, prior (11) is verified if likelihood of valid MIDI files is significantly greater than that of the corrupted MIDI files.

We have considered two types of corruption of A: (i) additive noise of increasing variance, and (ii) permutation of indices of active frames in the library. The latter corruption type is illustrated in Fig. 3. Likelihood of the corrupted amplitudes from the test set computed using parameters obtained from maximum likelihood approach on the training set are depicted in Fig. 4. The likelihoods are decreasing with increasing level of corruption. The decrease of likelihood with increased permutations is less significant than that with increased variance of the additive noise.

CONCLUSION AND FUTURE WORK

Two models of automatic music transcriptions were considered: model with strict restrictions on music events and model with relaxed restrictions. The process of estimation of the latter model is simpler, however, it does not penalize invalid sequences of music events. This penalization can be achieved by a suitable prior density function on parameters of the latter model. In this paper, we have made the first step in this direction

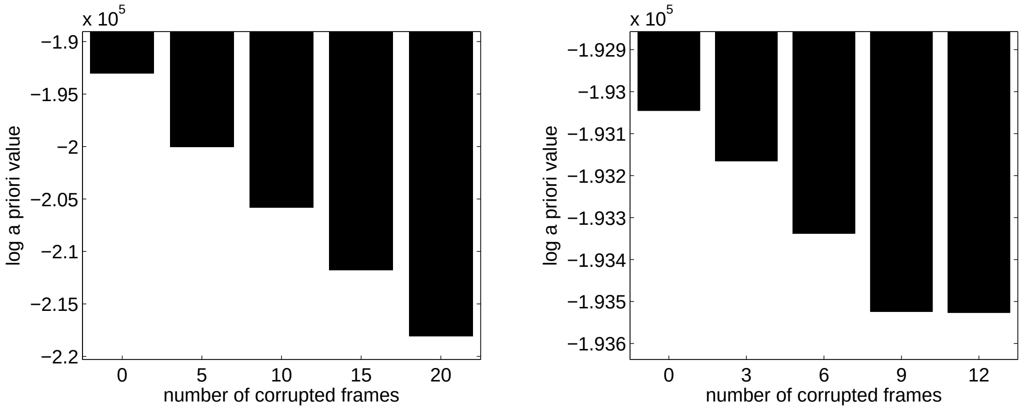

FIGURE 4. Value of prior for corrupted signals. Left: corruption by additive noise with random amplitudes of mean 0.5 and variance 0.0625. Each experiment was performed with different number of corrupted frames within each 20 time units (frames) of A. The likelihood of corrupted data is significantly decreasing. Right: permutation corruption for different number of corrupted frames per each 20 active frames of A.

by constructing the prior using methods of probability combination and minimization of Kullback-Leibler divergence. We have shown on synthetic data that the resulting prior density prefers valid music sequences over corrupted ones. Further improvement can be achieved via better optimization of the nuisance parameters, or by addition of other relevant pieces of prior knowledge. The presented approach is very general and can be immediately used in other application domains. The ultimate test of quality of the designed prior pdf is performance of the full music transcription algorithm.

REFERENCES

1. T. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley, 1958.
2. P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36:287–314, 1994.
3. M. Davy and C. Dubois. A fast particle filtering approach to bayesian tonal music transcription. 2007.
4. M. Davy, S. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *J Acoust Soc Am*, 119(4):2498–517, 2006.
5. M. Davy and A. Klapuri, editors. *Signal Processing Methods For Music Transcription*. Springer, 2006.
6. C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1:114–148, 1986.
7. H. Jeffreys. *Theory of Probability*. Oxford University Press, 3 edition, 1961.
8. M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2005.
9. K. Kashino and H. Tanaka. A sound source separation system with the ability of automatic tone modeling. In *International Computer Music Conference (ICMC)*, Aug. 1993.
10. S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–87, 1951.
11. S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.